Toward Transparent Optical Networking in Exascale Computers

Sébastien Rumley[1], David M. Calhoun[1], Arun Rodrigues[2], Simon Hammond[2], Keren Bergman[1]

[1]Columbia University, Sandia National Laboratories[2], bergman@ee.columbia.edu

Abstract *We review the requirements and expectations of future Supercomputer architectures, analyse how transparent optical networking might contribute to these future systems scalable performance, and develop the most immediate challenges for optical system developers in this context.*

Introduction

Over the past few decades, Supercomputers performance as signified by the “Top 500” list has steadily grown at an almost constant rate. Novel applications of high performance computing have emerged, from brain modelling to material design exploration, further adding to "traditional" applications ranging from multi-physics simulation to weather forecasting and climate modelling. More recently, the Big Data paradigm has further pushed computing needs toward processing of large volumes of data at ever increasing speeds. All these trends incite vendors to develop more powerful Supercomputers, soon poised to reach the mark of an ExaFLOP, i.e. 10^{18} floating-point operations per second.

To obtain such computing power, however, requires millions of processors (cores) to be gathered, interconnected and orchestrated. The realization and the operation of such computing systems pose serious challenges on various fronts. In particular, the interconnect architecture must support colossal amounts of data that is almost continuously inter-exchanged by the cores, as well as data flows present between the cores and the various memory resources.

Optical systems have proven to adeptly aid in high speed, high bandwidth data transfer over warehouse-scale distances. Due to the physical sizes of Supercomputers and large Data-centers, some links are long enough to justify a jump to optical domain (i.e. an E/O conversion at the emitter and an O/E conversion at the receiver). As optical technology further advances and the demand for bandwidth progresses, an increasing portion of the links present in Supercomputers (and, in the future, smaller scale computing platforms) will rely on optical interconnects. With this in mind, it is critical to examine the key system properties that would benefit by optically interconnected architectures. In this tutorial, current and forecasted Supercomputer architectures will be briefly presented, and future expectations in terms of photonic communication link features will be developed.

After this link level analysis, the tutorial will focus on transparent optical switching. Because network topologies included in large scale Supercomputers or Data-centers include tens of thousands of elements, it is common to see messages traveling over multiple hops, several of which may be optical. In this context, it is reasonable to investigate whether spatial optical switching can spare E/O and O/E conversions, and thus reduce the network power consumption footprint.

Computer architects are currently searching for technologies that can mitigate the negative consequences of extreme parallelization. Photonic interconnected networks can address these challenges. However, the lack of buffers in the optical domain introduces important limitations. Therefore, it is critical for the photonic community to appropriately examine the critical challenges for adoption of transparent optical networking, as well as the potential system performance advantages.

Extreme-scale architectural requirements

Most recent and upcoming Supercomputers are organized around tens of thousands of “nodes” (pictured in Fig. 1), each embedding one or more chip multi-processors (CMPs), a set RAM

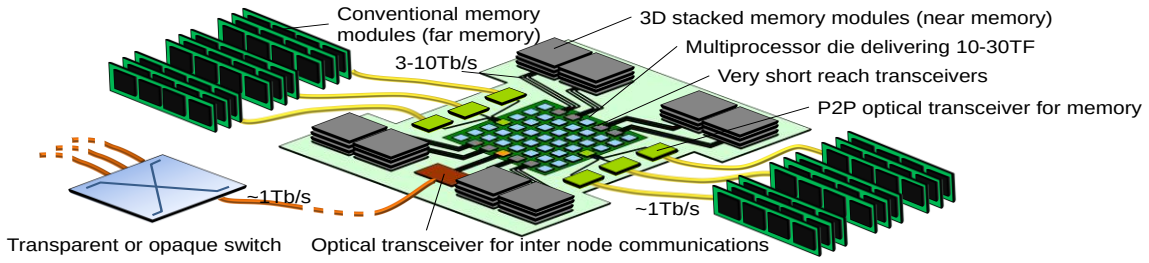

Fig. 1 Architectural organization of Exascale supercomputer node.

modules and one or more "I/O" interfaces (network, permanent storage). Nodes also increasingly include Graphical Processing Units (GPUs) computing chips that continue to populate video cards for nearly two decades.

The intrinsic computing power of a single core generally stays within tens of GF (GigaFLOP/s). However, the total power of a chip can reach several TF (TeraFLOP/s) by aggregating approximately 100 cores. Intel's recent Knights Landing" architecture collects 60 cores for a total computing power larger than 3TF. As it is possible to place several of them inside a node, aggregated computing power of *10s of TF* can be expected at the node level by the 2020 horizon.

To be appropriately exploited, all cores need to be constantly nourished with fresh data to process while simultaneously offloading results. The bandwidth necessary for these operations is proportional to the compute power. On the memory side, each FLOP (floating point operation) induces a typical displacement of a few bits (2 to 4) to and from the memory. This translates to *total memory bandwidth per node reaching the 100Tb/s mark* (e.g. 40TF x 4 bit/FLOP = 160 Tb/s). Since no single memory module will offer this bandwidth in the near-term, both memory capacity and bandwidth will be distributed over independent modules and links. Each FLOP must also be matched with a network bandwidth of 0.1 to 0.4 bits/s. Nodes' network interfaces are therefore expected to provide bandwidths at the Tb/s scale.

While both ranging in the 1-10 Tb/s region, memory and network links can be expected to exhibit different utilization. Relatively high utilization rates (>50%) are forecasted for memory links, the cores being constantly obliged to update their working material (this assumes that cores and memory modules are adequately mapped to each other). By contrast, lower average utilization rate of network links (<30%) have been measured. Cores tend to exchange data sporadically, with short periods of activity followed by relatively long periods of inactivity.

Besides the obvious procurement costs considerations, transmission systems in future Supercomputers will increasingly be subject to energy consumption constraints. Supercomputers already consume Megawatts (17MW for the most powerful systems). Next to the energy cost, this also represents a substantial challenge for the electric grid; beyond 100MW, a dedicated power plant might be required[1]. Taking 100MW as ceiling power and provided that only around 2/3 of the power is applied to the compute system (the remainder being consumed by cooling, power transformers, etc.), the power budget is *66MW at most* which corresponds to *66pJ/FLOP*. A FLOP involves at least 100 bits (two input operands, one result, one instruction, 4 bytes each = 128 bits), so the maximum energy budget per bit is *0.66pJ/bit*.

This mark provides a critical point of reference for the network requirements. If 1/1000 of the bits use the network (equivalent to 0.1 bit/FLOP), and shifting a bit over the network dissipates on average 100pJ, the network consumption already represents 15% of the envelope. There is no well-established limit indicating the ideal share of network consumption. Nevertheless, this indicates that 100pJ/bit is the far upper-limit for network operations, especially knowing that parallelization tends to increase the "verbosity" (i.e. higher bit-per-FLOP figures).

With 100pJ/bit as a metric encompassing the whole end-to-end network operation – including initial and final processing – one can extrapolate an approximate 20pJ/bit power budget per hop, since topologies are likely to involve 2-4 hops on average[2]. One switching operation is included in this extrapolation. Similar analyses will be provided for memory systems. These metrics become more constraining if higher performance is required, for example a network scalable to deliver 1.0 bit/FLOP.

Technologies for point-to-point transmission

The above analyses indicate the need for links capable of carrying at least 1Tb/s while limiting the energy dissipation per bit to 5-10pJ/bit, thus allowing 10-15 pJ/bit for switching. Expressed slightly differently, the power consumption should be limited to *5-10mW/Gb/s* at full load. A single 10Gbps data-carrying wavelength should therefore not require more than 100mW. As a point of comparison, commercially available active optical cables (AOC) can provide 40Gb/s for 0.78W (so 19.5mW/Gb/s)[3].

More integrated transceivers are necessary to realize these figures. It will be shown how, based on recent experiments involving silicon photonic links, power consumption below the 1mW/Gb/s can be predicted.

Toward transparent switching

It is always worth investigating transparent switching for the scenarios where multiple hops are required. Major challenges associated with transparent optical switching in a high-speed environment can be separated into three areas.

Power: Optical losses and perturbations (power penalties) are incurred over the optical path, which must be compensated by increasing the power level of an optical signal at its origin, or by optical amplification along the signal path. These

measures add to the overall power consumption. A challenge resides therefore in building highly power-penalty optimized switches. In parallel, lasers and amplifiers with high wall-plug energy efficiencies should also be targeted.

Optical path unavailability: During the time an optical switch adapts its physical state to change the signal propagation direction – subsequently inhibiting a sufficient optical power level on both switched signal paths – no usable data signal can be transmitted. Following this break in the signal, several link re-initialization mechanisms must take place, extending the time during which data cannot be transmitted. This can severely impact the overall switch throughput, especially if frequent re-adaptations are necessary. Physical "switching times" should therefore be optimized, but minimizing link re-synchronization times is equally important[4].

Bufferless arbitration: As of now, no effective random access buffering techniques are available in the optical domain. Therefore, optical switches cannot arbitrarily realign payloads over time to solve contentions. This constraint calls for additional hardware means to allow the optical switches and the network end-points to continuously synchronize their future allocations plans. Ultra-fast heuristics capable of minimizing both optical switch reconfiguration and payload latencies, as well as maximizing network throughput are also required.

Each of these challenges will be examined, as well as the most recent research progresses and future directions toward overcoming them.

Showstoppers

A major challenge for optical links is their general non-energy proportionality: energy consumption is almost independent of the number of bits effectively carried. This is primarily attributed to the continuous operation of the source lasers and photodetectors. Thus, if the links are not heavily utilized, much energy can be wasted in traffic scenarios where high bandwidths are needed, but only for short periods of time. Very good proportionality has been recently reported on electrical links[5], and similar efforts are required for optical links.

With extreme parallelism, quality-of-service (QoS) becomes an increasingly important aspect due to costly measures involved to checkpoint, restore and resume workloads after an error. Transmission bit-error rates should therefore be maintained below 10^{-12}. Errors can also be caused by irremediable packet losses, typically following a faulty or incomplete link initialization, potentially due to optical power transients. Obtaining guaranteed QoS can thus contradict fast link synchronization objectives, especially in the context of optical switching.

In the context of "dusty" traffic, composed of small bits of data destined to a multitude of diverse destinations, optical switches will quasi-perpetually adapt their states if setup/switching times cannot be reduced. An alternative is to use a conventional electronic network in parallel to carry the smallest messages, but this raises additional challenges (e.g. how to distinguish between "electrical packets" and "optical" ones).

Integrating nanophotonic components at the hardware level is one important achievement. However, it must be followed by efforts for developing interfaces between the photonics hardware, and the multiple computing and memory node elements. Dedicated experimental research towards system-level implementation is required to achieve efficient exploitation of photonics enabled architectures in practical environments[4].

Finally, as indicated above, electrical router chips are expected to consume 10-15 pJ/bit while offering high flexibility. To compensate for this lack of flexibility, optical switches will have to drive performance figures to significantly lower energies, reaching close to the 1pJ/bit (i.e. 1mW/Gb/s) mark. Even with state-of-the-art ring resonators, the tuning and trimming of rings alone consumes more than this value[6]. Further optimization efforts are therefore necessary.

Conclusions

Optical systems have made significant progress thanks to integration efforts, and they now populate an increasing number of systems. Various additional efforts are required to achieve ubiquity, however, especially towards reaching fully transparent optical interconnected systems.

References

[1] A. Marathe et al., "A Run-Time System for Power-Constrained HPC Applications," ISC, Frankfurt, (2015).

[2] S. Rumley et al., "Design Methodology for Optimizing Optical Interconnection Networks in High Performance Systems" Proc. ISC, Frankfurt (2015).

[3] M. Zuffada, "The industrialization of the Silicon Photonics: Technology road map and applications", Proc. ESSDERC, Bordeaux, (2012).

[4] D. M. Calhoun et al. "Integrating silicon photonics interconnects with computing systems", in preparation.

[5] W.-S. Choi et al., "A Burst-Mode Digital Receiver With Programmable Input Jitter Filtering for Energy Proportional Links" IEEE JSSC, Vol. **50**, no. 3, (2015).

[6] D. Nikolova et al., "Scaling silicon photonic switch fabrics for data center interconnection networks" Optics Express, Vol. **23**, no. 2, (2015).